DOCUMENT RESUME

ED 084 309 TM 003 333

AUTHOR	Canahl, Kathryn D.
TITLE	Psychological Scaling In A Public School Speech Therapy Program. Final Report.
INSTITUTION	Emory Univ., Atlanta, Ga. Div. of Allied Health Professions.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C. Regional Research Program.
BUREAU NO	BR-2-D-043
PUB DATE	1 Oct 73
GRANT	OEG-4-72-0021
NOTE	30p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Articulation (Speech); Child Language; Language Research; Rating Scales; Reliability; Sampling; *Speech Evaluation; *Speech Handicapped; *Speech Tests
IDENTIFIERS	*Scaling

ABSTRACT

The present study evaluated methods by which speech samples for equal appearing interval scaling of articulation defectiveness could be prepared with a minimum time expenditure. Ten public school speech therapists rated 40 one minute edited and 40 unedited speech samples on a nine point equal-appearing interval scale. The speech samples were recordings of the conversational speech of 40 children, classified as having articulation defects, ranging in age from 6-1[illegible]. The results indicated: (1) the average ratings for the unedited tapes is as reliable as the average rating for the edited tapes. (2) The agreement between judges and the mean scale values are essentially the same for the edited and unedited tapes. (3) There is no relation between time to make a rating and reliability. It was concluded that since the amount of preparation time for edited tapes is so much greater than for unedited tapes, that unedited speech samples are the more practical clinical tool and the ratings obtained will be as reliable as ratings for edited samples. (Author)

ED 084309

NOV 2 1 1973

Final Report

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: TM CS
In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

Project No. 2-D-043
Grant No. OEG-4-72-0021

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Kathryn D. Canahl, Ph.D.
Emory University
School of Medicine
Division of Allied Health Professions
Atlanta, Georgia 30322

TM 003 336

PSYCHOLOGICAL SCALING IN A PUBLIC SCHOOL SPEECH THERAPY PROGRAM

October 1, 1973

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Office of Education

Division of Educational Research

Region IV

FILMED FROM BEST AVAILABLE COPY

ABSTRACT

The present study evaluated methods by which speech samples for equal-appearing interval scaling of articulation defectiveness could be prepared with a minimum time expenditure. Ten public school speech therapists rated 40 one minute edited and 40 unedited speech samples on a nine point equal-appearing interval scale. The speech samples were recordings of the conversational speech of 40 children, classified as having articulation defects, ranging in age from 6-12. The results indicated: (1) the average ratings for the unedited tapes is as reliable as the average rating for the edited tapes. (2) the agreement between judges and the mean scale values are essentially the same for the edited and unedited tapes. (3) there is no relation between time to make a rating and reliability. It was concluded that since the amount of preparation time for edited tapes is so much greater than for unedited tapes, that unedited speech samples are the more practical clinical tool and the ratings obtained will be as reliable as ratings for edited samples.

Final Report

Project No. 2-D-043
Grant No. OEG-4-72-0021

Psychological Scaling in a Public School Speech Therapy Program

Kathryn D. Canahl, Ph.D.
Emory University
School of Medicine
Division of Allied Health Professions
Atlanta, Georgia 30322

October 1, 1973

TM 003 023

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgement in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Division of Educational Research
Region IV

ACKNOWLEDGEMENTS

The author wishes to express her appreciation to Mrs. Adelaide Beall and the DeKalb County School System Speech Therapists who provided their time during the data collection portion of the present study.

Special appreciation is extended to Julius A. Canahl, Ph.D. Research Scientist Veterans Administration Hospital, Atlanta, Georgia for invaluable assistance in data analysis.

TABLE OF CONTENTS

Chapter		Page
	Acknowledgements	iv
	Table of Figures	vi
I	Introduction	1
	Statement of Purpose	1
	Introduction	1
	Review of Literature	3
	Scope of the Problem	4
	Specific Problem	4
II	Procedure	6
	Preliminary Study	6
	Procedure for Preliminary Study	6
	Main Study Part I	7
	Procedure	7
	Main Study Part II	8
	Procedure	8
III	Results	10
	Scale Values	10
	Reliability of Average Ratings	10
	Reliability of Individual Ratings	10
	Relationship Between Ratings for Edited and Unedited Samples	10
	Relationship Between Time to Make a Rating and Reliability	10
IV	Discussion and Conclusions	16
	Discussion	16
	Conclusions	17
V	Recommendations	18
Bibliography		19
Appendix A		23

TABLE OF FIGURES

Figure		Page
1	Histogram showing the distribution of time required to make a judgment for edited samples.	12
2	Histogram showing the distribution of time required to make a judgment for unedited samples.	13
3	Scatter point graph showing the relation between the mean time required to make a rating and reliability of the rating for edited samples.	14
4	Scatter point graph showing the relation between the mean time required to make a rating and reliability of the rating for unedited samples.	15

CHAPTER I

INTRODUCTION

Statement of Purpose. Psychological scaling has been proven to be a valuable research technique in speech pathology. The specific procedures by which this technique can be applied to the assessment of communicative disorders outside of a laboratory setting have yet to be demonstrated. The purpose of the present project was the evaluation of methods by which the speech samples used in scaling could be prepared with a minimum time expenditure. If preparation time can be reduced sufficiently without loss of accuracy, the use of scaling in a public school speech therapy program could result in major increases in program efficiency.

Introduction. Although there are many parameters of speech, articulation seems to play a major role in the intelligibility of speech and therefore oral communication. Articulation may be defined as "the acoustic impression, the distinctness, or acceptability of the speech sound" (42). An articulation error is defined as the defective production of a specific phoneme which may lead to a deficit in intelligibility. Mills and Streit (18) report that 7% of the overall school population have a defect in articulation. The 1961 ASHA Survey (12) reports that approximately 81% of the public school therapist's case load is made up of children with articulatory problems.

One of the methods most frequently used to identify those children with articulation problems is the speech survey (12). In the speech survey method, the speech therapist administers a screening test such as the Templin Darley 50 items screening test (39). Although this type of test may provide an inventory of misarticulated sounds it often does not offer a complete picture of the effect of the articulation defect on communication. It is known that many children are variable in the type of misarticulation they produce. In one instance the child may omit a phoneme in a word, in another sample of speech he may distort or substitute for the previously omitted phoneme. In each situation the specific type of misarticulation will have a different effect on the intelligibility of the particular word.

In view of the above, it has been suggested by Johnson, Darley, and Spriestersbach (10) that in addition to an analysis of errors, the speech clinician utilize a four point scale to rate the overall speech behavior of an individual. The various categories for the four point scale are defined as follows: 1.) adequate, within the normal range; 2.) a deviation which does not make for a communicative handi-

cap; 3.) a deviation which occasions a moderate communicative handicap that should be given clinical attention; and 4.) a deviation which results in a serious communicative handicap that requires immediate clinical attention" (Johnson et al 10). While this is an attempt to obtain information regarding the severity of the communication defect, there are several apparent weaknesses in this approach. In the first place, the type of rating suggested by Johnson et al indicates that the rater and the subject are face to face. Research has indicated, however, that inter-examiner reliability is higher when judgments are made from recordings (0.71-0.85) than when the judgments are made in a live condition (0.61-0.75) (45). Intra-examiner reliability also appears to be higher when judgments are made from recordings. Wright (45) reported that self agreement between live and recorded judgments is lower (0.55-0.75) than those obtained when all judgments are made from recordings (0.76-0.80). Secondly, although this rating method may prove satisfactory to a specific clinician in a specific situation, it has not been demonstrated that what one clinician labels as a 3 will be labeled as a 3 by another clinician. If they do not agree, the ratings have little value beyond the individual clinician's own use of his data.

In addition, the public school speech clinician is usually faced with the problem of having the number of children who need help far exceeding the number she can adequately handle. Using a four point scale in which only two categories are used for cases needing clinical attention often will not solve the problem of case selection for the therapist. It is conceivable that the therapist might have more children rated a 4 ("a deviation which results in a serious communicative handicap that requires immediate clinical attention") than she can handle. Yet, this rating method does not help her to decide which 4's she should take. It would seem that more information could be provided if the number of categories were enlarged to allow more refinement. For example, if the clinician used a scale with nine categories, she could utilize ratings of 1, 2, 3 for cases with no or minimal problem; 4,5,6 for cases with moderate severity and 7,8,9 for the very severe cases. Finally, the rating method suggested by Johnson et al provides no information as to how much more severe the subject rated 4 is compared to the subject rated a 3; only that he is more severe.

The rating scale suggested by Johnson et al is only one of many scaling techniques that could be utilized in the area of speech defectiveness. One psychological scaling method that has been used extensively in speech pathology research is the method of equal-appearing intervals. In utilizing this method the judge is instructed to assign numbers to the stimuli in relation to an equal appearing interval scale of severity. One of the underlying assumptions of this method is that the judge is capable of sorting the stimuli on an equal-interval subjective scale (40). If we can accept this assumption, then we have some indication as to how much more severe a 4 is than a 3 or an 8 is than a 7. Since the method has no provision for testing this basic assumption (40), previous investigators (e.g., 14, 20, 32) have been willing to accept this assumption when the reliabilities have been high (0.89-0.97). In further support of this assumption, research has shown that there is a relation-

ship between the scale value assigned a child and the number of misarticulated sounds as measured by a phonetic inventory. Jordan (11) investigated the relationship between articulation test measures and listener ratings of articulation defectiveness. He reports that the measures correlating most highly with judged severity were (1) the number of defective items (Pearson r. 0.72), (2) the number of defective sounds (Pearson r. 0.75) and (3) the number of defective singles (Pearson r. 0.78). In summary, although the assumption that the subject is capable of sorting the stimuli on an equal-interval subjective scale is not directly testable, it would appear that because of the high average reliabilities, the good agreement between judges and the relationship between scale values and other measures of articulation defectiveness, psychological scaling in this area would seem to be valid.

Review of Literature. Reliable scale values have been obtained using the equal appearing intervals method in the following areas of speech defectiveness: articulation (11, 20, 29, 31, 32); cleft palate (6, 9, 15, 19, 24, 25, 28, 37, 38, 41); language (33, 34, 35, 36); stuttering (1, 5, 13, 14, 17, 23, 26, 30, 43, 46, 47); and voice disorders (4, 21, 22, 27). Since the present study is concerned specifically with articulation defectiveness, further discussion will be directed to this particular problem.

Reliable scale values of articulation defectiveness have been obtained by the methods of equal-appearing intervals, paired comparisons, and successive intervals (31). The results of the Sherman and Moodie (31) study indicated that the equal-appearing intervals method was most useful in scaling articulation. Morrison (20) investigated the reliability of scale values when judges rated 5 and 10 second segments of one minute continuous speech samples of children. Her results indicated that reliable scale values could be obtained for both the 5 and 10 second segments. The Pearson r's obtained were 0.97 for the 5 second samples and 0.96 for the 10 second samples. Sherman and Morrison (32) investigated the reliability of individual ratings of severity and from their results conclude that reliable mean scale values can be obtained from the responses of a trained individual judge using one minute speech samples (Pearson r's: 0.93-0.99).

In the previous studies of Morrison (20) and Sherman and Morrison (32), the judges rated one minute speech samples in segments of 5 or 10 seconds. In other words, the one minute samples were broken down into small segments and tapes were constructed in which each individual's 12 (5 second) or 6(10 second) segments were randomly presented. Sherman and Cullinan (29) investigated the reliability of individual judges' ratings of articulation defectiveness based upon single ratings of one minute speech samples. They report that reliable mean scale values were obtained from one minute samples as a whole (Pearson r's: 0.91-0.97). These results indicate that it is not necessary to break the one minute speech samples into five or ten second segments.

All of the previous studies used extensive training procedures. These training procedures were all similar and were based on the findings of Morrison (20) and Sherman and Morrison (32). Canahl and

Strumph (3) investigated if reliable ratings of articulation defectiveness of children could be obtained using therapists who were untrained in scaling as judges. Their results indicated that reliable scale values could be obtained without training the therapist in the specific area of scaling procedures (Pearson r's: 0.88-0.94).

In summary it has been demonstrated that reliable ratings of articulation defectiveness can be obtained using the method of equal appearing intervals (11, 20, 29, 31, 32) and these ratings correlate very well with the results of a phonetic inventory (11).

Scope of the Problem. It would seem that psychological scaling would be a very useful tool in many areas of the public school speech therapy program. Based upon the results of all the previous studies it would appear that a clinician might use the scaling method of equal appearing intervals as a screening device rather than a phonetic inventory or as a supplement to her phonetic inventory. First of all, the sample of conversational speech which is utilized in scaling is much closer to the "real life situation" and provides information on how the defect actually affects communication. Secondly, the use of scaling might save the clinician time. The purpose of a speech survey or screening program is to identify those children who need help. Rather than doing a time consuming inventory of articualtion errors, which is necessary for a complete diagnosis on all children, the therapist would need to do it only on those children whose ratings exceeded a certain scale value; for example 4.

The advantages of using a scaling technique are not limited to providing the therapist with a more realistic indication of the extent of the individual's communication problem. In many larger school systems, computers are being utilized to store and process infomation concerning individual students. It would appear that scaling would be an ideal way to provide one number to indicate the severity of a particular child's speech defect. This would enable the system to rapidly identify the speech therapy needs for a particular school within the system by simply counting the number of students falling within a particular category, e.g., "most severe", and as a result allow for more efficient utilization of available speech therapy personnel.

Specific Problem. Although scaling would appear to be a very useful technique in a public school speech program, scaling thus far has only been used as a research tool. It would appear that the reason scaling has not been used in the public school is that no one has investigated the basic question as to whether or not scaling would be practical in the clinical situation. It has been assumed that it would be impractical for the clinician since the method for collecting the speech samples to be scaled involves a considerable amount of time. In all of the previous studies the method used was as follows: The conversational speech of a number of children is tape recorded and the experimenter edits the tape so that each child's sample consists of one minute of continuous speech. This means that all pauses have been spliced out. Since the speech to pause time ratio of children is highly variable during conversational speech, a sample of continuous speech lasting one minute may require from five

to thirty minutes of unedited conversational speech. It is obvious that the amount of time necessary to prepare the tape for scaling in this manner (roughly 15 minutes to two hours) is much more than a public school speech clinician can afford. If scaling is to be a tool that can be used in the clinic as well as in research, the first question to be answered is how can one minimize the amount of time required for preparing the speech samples to be scaled.

The method used in all previous studies has two underlying assumptions: (1), the speech sample must be one minute in length and (2), the speech sample must be continuous speech, i.e., no pauses. Thus far, no one has attempted to investigate the validity of these assumptions. The purpose of the present study was to attempt to ascertain whether or not the above mentioned assumptions are valid. Specifically, the project was designed to answer the following questions:

1) How long does a continuous speech sample have to be in order for judges to reliably scale articulation defectiveness?

2) Can reliable ratings of articulation defectiveness be obtained using unedited conversational speech?

CHAPTER II

PROCEDURE

Preliminary Study. In order to insure that the speech samples for the main study represented a wide range of articulation ability, from normal to severely defective, a preliminary study was conducted. The original samples, recorded by the experimenter were rated by 4 speech pathologists who hold the Certificate of Clinical Competency from the American Speech and Hearing Association.

Procedure for Preliminary Study

Stimulus Material. Recordings of the conversational speech of 50 children were obtained. The children all came from the DeKalb County school system, DeKalb County, Georgia and ranged in age from 6 through 12. Each child was evaluated by the school speech therapist as having an articulation defect. The recordings were made in quiet conditions using a Sony 850-2 tape recorder. The tape speed was 7 1/2 inches per second.

Copies of the original sample were made by dubbing from a Sony 850-2 tape recorder to a second Sony 850-2 tape recorder. This was done to preserve the original unedited tape and to allow construction of an edited tape recording with one minute continuous speech samples from each of the 50 children. One minute was chosen on the basis of the results reported by Sherman and Cullinan (29) which indicated that reliable scale rules could be obtained for ratings of articulation defectiveness using one minute speech samples. On the final edited tape each child's one minute speech sample appeared twice in random order. Immediately preceding each sample a number was recorded to identify the sample to the judges. At the end of each sample, the experimenter inserted the word "rate" to instruct the judges to make their ratings. Between each sample there was a pause of five seconds during which the judges recorded their ratings.

Judges. March, Weaver, Morrison and Black (16) reported that reliable average scale values could be obtained with a panel of four or more judges. For this reason four speech pathologists who hold the Certificate of Clinical Competency by the American Speech and Hearing Association served as judges.

Listening Conditions. The judges listened to the recorded speech samples in a quiet classroom. Based upon the information from previous studies (20, 31, 32, 16, 29), the judges rated each child on a nine point equal-appearing interval scale with 1 representing least severe and 9 most severe articulation defectiveness.

Selection of Samples for Main Study. The intraclass correlation

technique for evaluating the reliability of mean scale values as described by Ebel (8) was applied to the data obtained. The value obtained for average ratings was 0.94. This reliability estimate compares favorably with the intraclass correlation coefficient for average ratings reported by previous investigators. The range of average reliabilities reported in previous studies using the Ebel method is 0.89-0.97.

The reliability of each of the four observers was evaluated by correlating each observer's ratings with the mean ratings of the remaining three observers. The correlations obtained ranged from 0.80 to 0.91 (obtained r's 0.80, 0.88, 0.90, 0.91). The repeat reliability for each of the four judges was also obtained by correlating the ratings each judge made the first time he heard a specific child's speech sample with the rating he made the second time he heard the same child's speech sample. The correlations obtained ranged from 0.73 to 0.93 (obtained r's 0.73, 0.89, 0.84, 0.93).

All of the previous information and the range of mean scale values for each of the 50 samples was used to select forty samples to be used in the main study. The criterion for selection of the specific samples to be used in the main study was the agreement between judges. Samples which showed the least variation in scale values were chosen. In addition the forty samples were selected so that there was a range from normal to severely defective articulation.

Training Tape. A training tape was made to be used in the main study. The method used to construct the training tape was similar to the one previously employed by Morrison (20). Nine speech samples, each sample representing a specific level of severity from 1-9, were chosen from the 40 samples selected for the main study. The choice of the specific samples was based on the agreement between the judges. Two training tapes were constructed by dubbing from the edited tape used in the preliminary study. One training tape consisted of 30 second samples from each of the nine speech samples chosen. The second training tape consisted of 10 second speech samples from each of the nine samples chosen.

Main Study Part I. The purpose of this phase of the main study was to determine the mean length of time required for continuous speech samples to be scaled reliably for articulation defectiveness.

Procedure. Stimulus Material. The stimulus material consisted of the 40 one minute speech samples chosen on the basis of the preliminary study.

Judges. Ten speech therapists who were working as public school speech therapists in the DeKalb County Schools and who had not provided any children for the samples to be rated served as judges.

Training Session. A training session was held one half hour prior to the experimental listening session. Each judge was run indi-

vidually. The training tape with the nine 30 second speech samples, arranged in order of severity from least severe to most severe, was played to the judge. He was instructed to listen to the tape to obtain an idea of level of severity. The same nine speech samples, this time arranged in random order were then played to the judge. The judge was instructed to rate each sample on a nine point equal-appearing interval scale with one representing least severe and nine representing most severe. Following this the judges again listened to the nine 30 second speech samples presented in the same random order as in the previous listening session. During this listening the judge was informed of the correctness of his ratings. After completion of this task the whole procedure was repeated using the training tape with 10 second speech samples.

Experimental Listening Session. The experimental listening session immediately followed the training session. The judge listened to the tape of the forty one-minute speech samples. He was reinstructed to rate each sample on a nine point equal-appearing interval scale. In addition, the judge was instructed to make a rating as soon as he felt that he had heard enough of the speech sample. The experimenter, using a stop watch, recorded the time the judge took to make a rating. This procedure was followed until all 40 samples were rated.

Part II Main Study. The purpose of this phase of the main study was to evaluate the reliability of ratings made on the basis of unedited conversational speech samples. This phase of the main study was conducted one month after the completion of Part I.

Stimulus Material. The stimulus material consisted of the 40 original unedited tape recordings of conversational speech of children between the ages of 6 and 12. These recordings were the unedited versions of the 40 one minute speech samples used in Part I of the main study. In other words, the samples used were the same as those used in Part I except these recordings included all pauses, speech of the investigator, etc.

Judges. The ten public school speech therapists who served as judges for Part I of the main study also served as judges for Part II. This was done to allow a direct comparison between ratings of edited and unedited speech samples.

Training Session. One month after the judging session for Part I of the main study the judges were individually retrained using the same tapes and procedures described in Part I.

Experimental Listening Session. Immediately following the training session a judge listened to the experimental tapes. He was reinstructed on rating each sample on a nine point equal-appearing interval scale and to stop the tape recorder as soon as he felt he was confident enough to make a rating. The experimenter, using a stop watch, recorded the time it took to make a rating while the judge

recorded his rating. This procedure was followed until all 40 unedited samples were judged.

CHAPTER III

RESULTS

The main study, as you will recall, was divided into two phases. The first phase had ten judges rate continuous speech samples which will be called edited samples throughout the remainder of this report. The second phase had the same ten judges rate conversational speech samples which will be referred to as unedited samples.

Scale Values. Each judge rated each edited and unedited speech sample so that a total of 400 individual ratings were obtained for each condition, i.e. edited and unedited. The mean of the 10 ratings for each edited speech sample was calculated for a total of forty mean scale values. The same procedure was followed for the unedited samples. The individual ratings for both edited and unedited speech samples ranged from one to nine on the severity scale. A rating of 1 represented least severe and 9 most severe articulation defectiveness. The mean ratings for the forty edited and the forty unedited samples also ranged from 1-9.

Reliability of Average Ratings. The reliability of the mean ratings was evaluated by the intraclass correlation technique described by Ebel (8). The values obtained for the reliability of the mean rating for the edited samples was 0.96 and 0.97 for the unedited samples.

Reliability of Individual Ratings. The reliability of each of the ten judges was evaluated for both the edited samples and the unedited samples separately. This was done by correlating each judge's ratings, for a specific condition, with the mean rating of the remaining nine judges for the same condition. The Pearson r's obtained for the edited samples ranged from 0.73 to 0.91(r's obtained were 0.73, 0.80, 0.84, 0.84, 0.84, 0.84, 0.87, 0.88, 0.91). For the unedited samples the Pearson r's ranged from 0.79 to 0.94 (r's obtained were 0.79, 0.80, 0.83, 0.85, 0.86, 0.89, 0.90, 0.91, 0.91, 0.94).

Relationship Between Ratings for Edited and Unedited Samples. The correlation between the mean scale values for the 40 edited and 40 unedited speech samples was evaluated through the use of the Pearson r. The Pearson r obtained was 0.95.

Relationship Between Time to Make a Rating and Reliability. As previously stated the experimenter recorded the time it took each judge to rate each sample. From this information a mean time for each judge across samples was calculated. The range of mean times

for edited samples was from 8 to 39 seconds and the mean time for all samples across judges was 23 seconds. For the unedited samples the range was 21 to 35 seconds and the mean time for all samples across judges was 26 seconds.

Histograms shown in Figures 1 and 2 were also used to evaluate time. The frequency of occurence for the time each judge took for each edited sample is shown in Figure 1. The same graph for the unedited samples is shown in Figure 2. As can be seen in both figures, the distributions are positively skewed. The median score for the edited samples was 18 seconds; for the unedited 25 seconds.

Figures 3 and 4 show graphically the relation between the mean time for each judge to make a rating and his reliability evaluated by the Pearson r for the edited and unedited samples. As can be seen, there appears to be no linear relation, for either the edited or unedited conditions, between the length of time to make a rating and reliability.

Figure 3

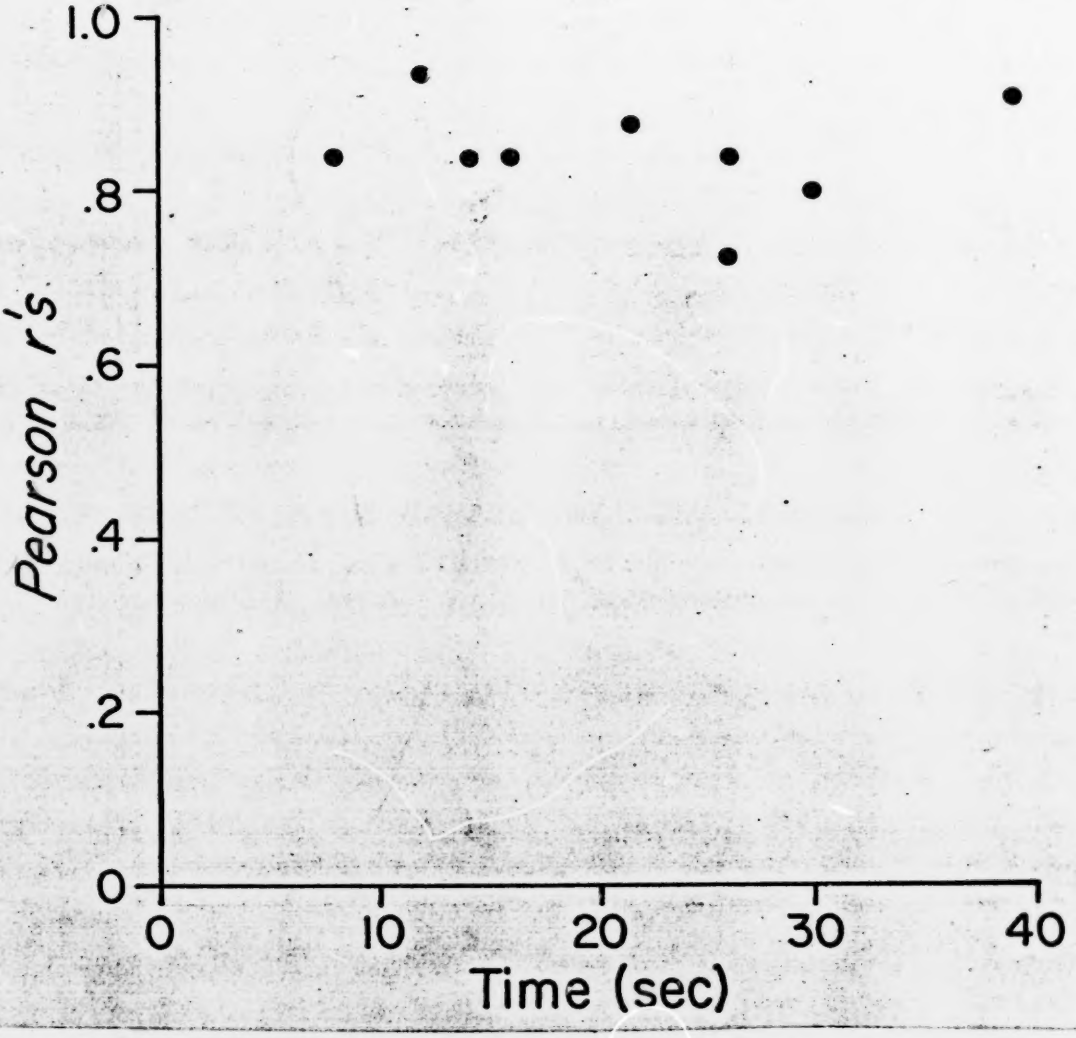

Fig. 3. Scatter point graph for edited samples showing the relation between the mean time required to make a rating and reliability of the rating. Each point represents one judge rating forty edited speech samples. The ordinate is the mean time required to make a rating. The absissa is the reliability as measured by the correlation between a specific judge's rating with the mean rating of the remaining nine judges for each speech sample.

Figure 4

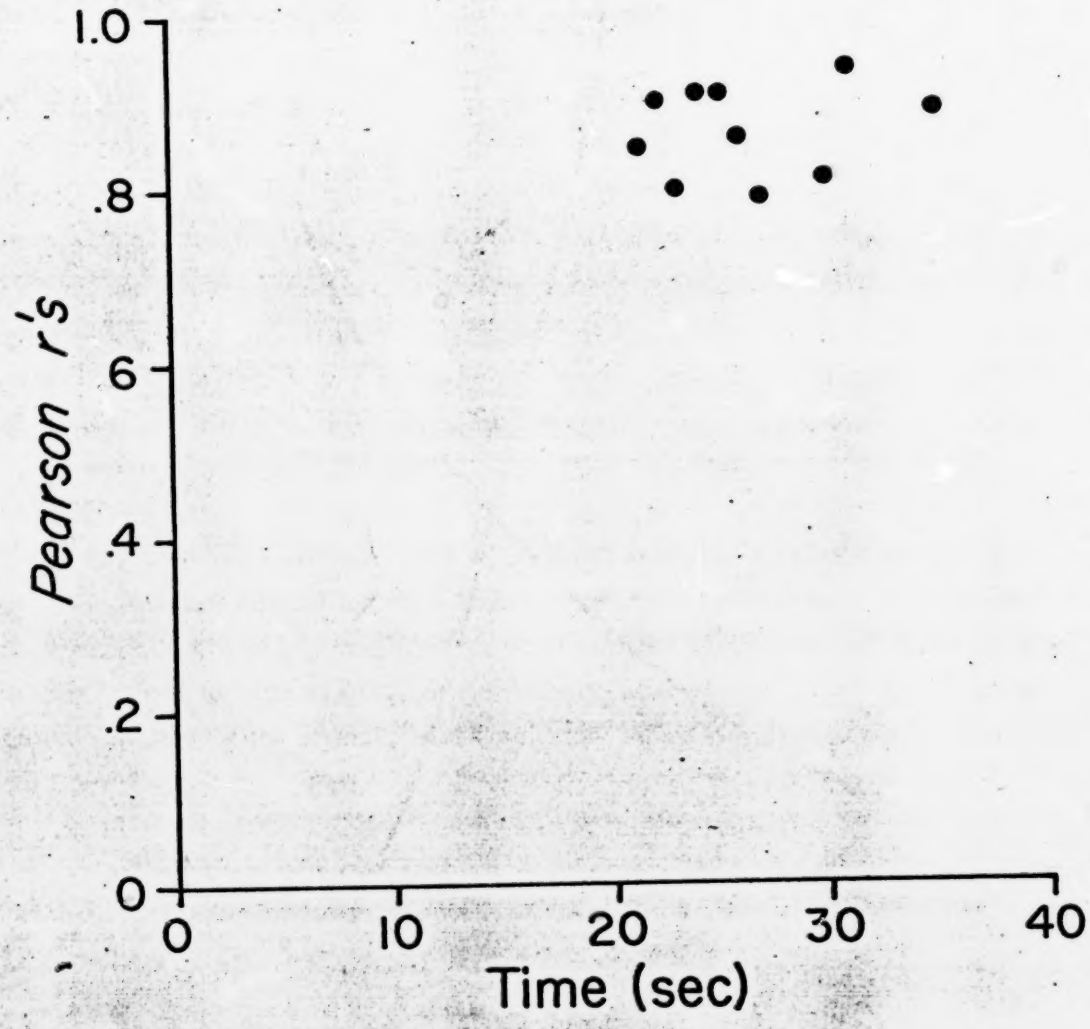

Fig. 4. Scatter point graph for unedited samples showing the relation between the mean time required to make a rating and reliability of the rating. Each point represents one judge rating forty unedited speech samples. The ordinate is the mean time required to make a rating. The absissa is the reliability as measured by the correlation between a specific judge's rating with the mean rating of the remaining nine judges for each speech sample.

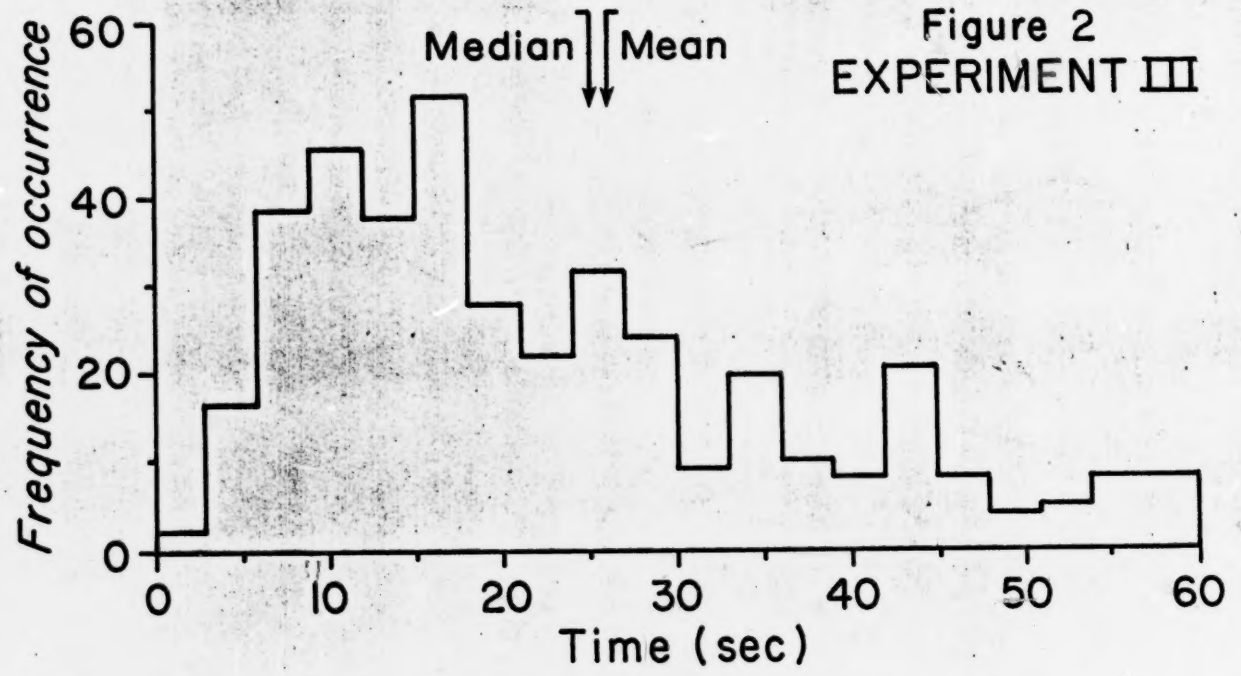

Fig. 2. Histogram showing the distribution of time required to make a judgment. Data is from ten judges each rating forty unedited speech samples.

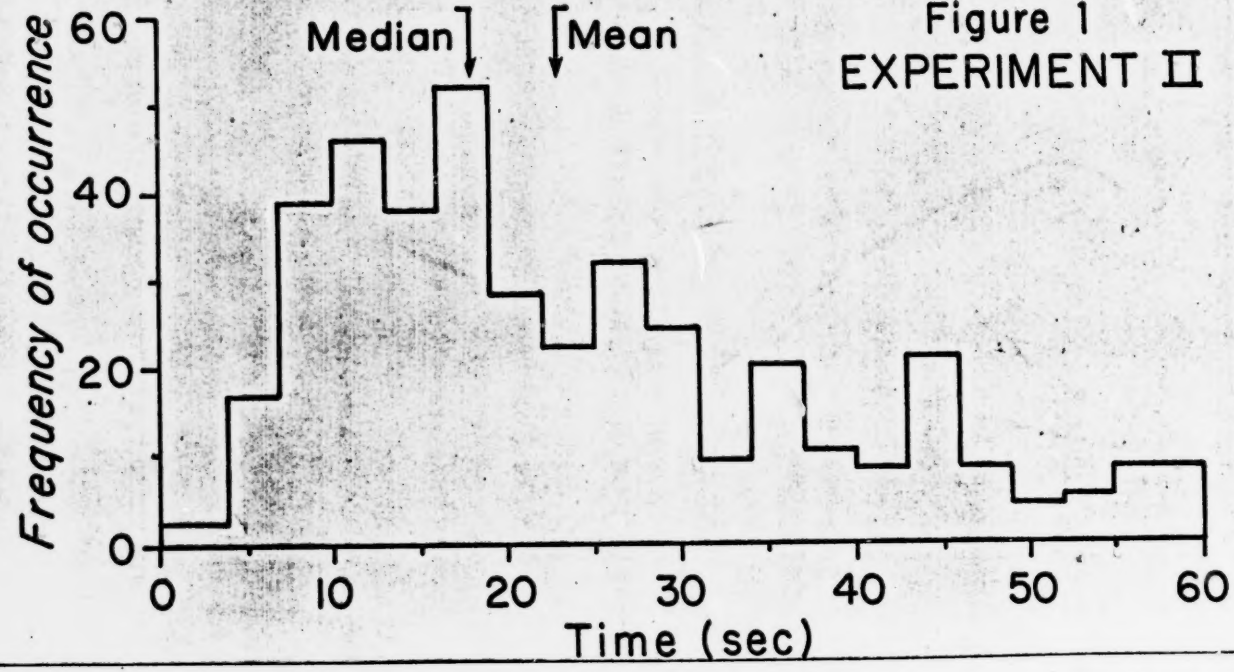

Fig. 1. Histogram showing the distribution of time required to make a judgment. Data is from ten judges each rating forty edited speech samples.

CHAPTER IV

DISCUSSION AND CONCLUSIONS

The purpose of the present study was the evaluation of methods by which the speech samples used in scaling could be prepared with a minimum time expenditure. Specifically, the project was designed to answer two questions. The first question was: How long does a continuous speech sample have to be in order for judges to reliably scale articulation defectiveness. The results of the present study indicate that there is little if any correlation between time and reliability within the confines of one minute samples. In previous studies the judges were all given speech samples of a specified, uniform time. In the present study each judge was allowed to determine for each sample how much time he needed to rate. A comparison of the reliability of the average rating obtained in the present study with the reliabilities obtained in previous studies indicates good agreement. The reliability coefficient obtained in the present study was 0.96. The range of reliabilities obtained in previous studies was between 0.89, as reported by Sherman and Cullinan (29), and 0.97 as obtained by March, Weaver, Morrison and Black (16). The agreement between the results of the present study and previous studies indicate that each judge is able to determine individually for each sample, the duration required to make a rating. Even though judges use different durations to make a rating, the average rating across judges will be reliable. As shown in Figure 2 even when you look at an individual judge there appears to be no linear relation between the length of time and reliability.

Another aspect of the relation between reliability and time to make a rating are the Pearson r's obtained for individual judges. The Pearson r's obtained in the present study for the edited speech samples ranged from 0.73 to 0.91. These correlations are low compared to those obtained in most previous studies. Canahl and Strumph (3) report correlations ranging from 0.88 to 0.94, and Sherman and Cullinan (29) report a range from 0.91 to 0.97. However, Bucher and Canahl, who also used public school therapists as judges, report a range from 0.79 to 0.92 which is in closer agreement with the present study.

The second question the present study was designed to answer was: whether or not ratings of articulation defectiveness could be obtained using unedited conversational speech. The results of the present study indicate that reliable ratings can be obtained using unedited speech samples. The reliability coefficient for the average rating of unedited samples obtained in the present study was 0.97. This agrees very well with the results of previous studies and the results obtained in the present study for edited samples. The range of reliabilities obtained in previous studies was between 0.89, as reported by Sherman and Cullinan (29), and 0.97 as obtained by March, Weaver, Morrison and Black (16). The reliability coefficient for the unedited

samples in the present study was 0.96. In addition to the agreement between the reliability coefficients obtained, the correlation between the mean scale values for the edited and unedited samples in the present study also demonstrated a close relation. The Pearson r obtained was 0.95.

The main purpose of the present study was to determine if the preparation time and possibly the listening time involved in scaling could be shortened so that scaling would become a practical tool for the public school speech clinician. The results of the present study demonstrate that unedited speech samples can be judged as reliably as edited samples. In addition, as stated previously, for the edited speech samples there is no linear relation between reliability and time to make a rating. Figure 4 shows that the same statement can be made with regards to the unedited samples. It would appear that although judges vary with respect to time to make a rating this has little if any relation to their reliability.

Another aspect is the amount of time a listener requires to rate edited samples as opposed to unedited samples. In the present study the mean time to make a rating for edited samples for the judges was 23 seconds. The mean time for the unedited samples was 26 seconds. This indicates that on the average a judge required three more seconds of listening time for each unedited sample. When one compares the amount of preparation time for the edited samples to the preparation time for the unedited samples the increase in listening time appears negligible. In order to prepare edited tapes the speech therapist must do the following: (1) record the samples (2) dub the original samples on to a second tape (3) splice out all pauses in the child's speech as well as any of her own speech. To prepare unedited tapes, for rating, all the therapist must do is record the child's speech. In view of this it would appear that the use of unedited speech samples for rating articulation is a much more practical tool.

Conclusions. The results of the present study indicate the following:

(1) the average rating for unedited tapes is as reliable (0.97) as the average rating for edited tapes (0.96).

(2) the agreement between judges for the present study, although not as high as in most previous studies, was essentially the same for the edited and unedited samples. The range of Pearson r's for the edited samples was 0.73 - 0.91 and for the unedited 0.79 - 0.94.

(3) there is good agreement between the mean scale values obtained for edited and unedited speech samples. The Pearson r obtained in the present study was 0.95.

(4) there is no linear relation between time to make a rating and reliability for either edited or unedited samples.

(5) the difference in listening time for edited and unedited samples is small. The mean time for edited samples is 23 seconds and for the unedited 26 seconds.

(6) the amount of preparation time for edited tapes is so much greater than for unedited tapes it is concluded that unedited speech samples are the more practical clinical tool and the ratings obtained will be as reliable as ratings for edited samples.

CHAPTER V

RECOMMENDATIONS

The results of the present study indicate that unedited speech samples can be utilized for rating articulation defectiveness. The average rating for the unedited samples was as reliable as the average rating for edited samples. In addition, the unedited samples require less time to prepare and require, on the average, a small amount (three seconds) of additional listening time. In view of this it would appear that equal-appearing interval scaling can be a useful tool in a public school speech therapy program. This type of scaling should be useful in the following areas: a. speech screening b. pupil progress c. therapists effectiveness d. training new therapists and e. training classroom teachers. Specific details on the possible utilization of equal-appearing interval scaling in each of these areas may be found in Appendix A.

In the present study the intraclass correlation coefficient for the average rating as well as the range of Pearson r's for individual judges was higher for the unedited samples. The reliability for the average rating for edited samples was 0.96 and 0.97 for unedited samples. The range of the Pearson r's for individual judges for the edited samples was 0.73 - 0.91 and for the unedited samples 0.79 - 0.94. In the present study all judges rated the edited samples and then one month later rated the unedited samples. It is possible that the reliabilities are higher for the unedited samples due to a learning effect. That is, the reliabilities for unedited samples may be higher because the judges had more experience in rating speech samples on an equal-appearing interval scale. This would seem to indicate the need for research in the area of training and its effect on rating. Specifically two questions should be answered.

(1) How much training prior to rating is necessary to obtain reliable ratings?

(2) Does training need to be repeated at various intervals in time in order to maintain a judge's reliability?

Even though it is suggested that research be done in the area of training, this does not negate the immediate use of scaling in public school speech therapy programs. The training period in the present study was short, a half hour, and yet the reliabilities obtained were in good agreement with previous studies.

References

1. Boehmler, R.M., Listener responses to non-fluencies. J. Speech Hearing Res., 1, 132-141 (1958).

2. Bucher, D., and Canahl, K., Comparison of ratings of defective articulation by a group of first grade teachers and a group of speech therapists. Unpublished study, 1971.

3. Canahl, K., and Strumpf, S., Scaling articulation defectiveness. Unpublished study, 1970.

4. Clarke, F.R., and Becker, R.W., Comparison of techniques for discriminating among talkers. J. Speech Hearing Research, 12, 747-761 (1969).

5. Cullinan, W.L., Prather, Elizabeth M., and Williams, D.E., Comparison of procedures for scaling severity of stuttering. J. Speech Hearing Res., 6, 187-194 (1963).

6. Dickson, D.R., An acoustical study of nasality. J. Speech Hearing Res., 5, 103-111 (1962).

7. Diehl, C., Sinnett, C., Efficiency of teacher referrals in a school speech testing program. J. Speech Hear. Dis., 24, 34-36 (1959).

8. Ebel, R., Estimation of the reliability of ratings. Psychometrika, 16, 407-424 (1951).

9. Hess, D.A., Pitch, intensity, and cleft palate voice quality. J. Speech Hear. Res., 2, 113-125 (1959).

10. Johnson, W., Darley, F., and Spriestersbach, D., Diagnostic Methods in Speech Pathology. New York, Harper and Row, 1963.

11. Jordon, E.P., Articulation test measures and listener ratings of articulation defectiveness. J. Speech Hear. Res., 3, 303-319 (1960).

12. Knight, H., Hahn, E., Ervin, J. and McIsaac, G., The public school clinician: Professional definition and relationships. J. of Speech Hear. Dis. Monogr. 8, 10-21 (1961).

13. Lanyon, R.I., The relationship of adaptation and consistency to improvement in stuttering therapy. J. Speech Hear. Res., 8, 263-269 (1965).

14. Lewis, D. and Sherman, D., Measuring the severity of stuttering. J. Speech Hear. Dis., 16, 320-326 (1951).

15. Lintz, L.B., and Sherman, D., Phonetic elements and perception of nasality. J. Speech Hear. Res., 4, 381-396 (1961).

16. March, N., Weaver, C., Morrison, S., and Black, J., Observed and predicted estimates or reliability of aspects of a speech articulation rating scale. Speech Monogr. 25, 296-304 (1958).

17. Martin, R., Stuttering and perseveration in children. J. Speech Hear. Res., 5, 332-339 (1962).

18. Mills, A.W., and Streit, H., Report of a speech survey, Holyoke, Massachusetts, J. Speech Dis., 7, 161-167 (1942).

19. Morris, H.L., Communication skills of children with cleft lips and palates. J. Speech Hear. Res., 5, 79-90 (1962).

20. Morrison, S., Measuring the severity of articulation defectiveness. J. Speech Hear. Dis., 20, 347-351 (1955).

21. Rees, M., Harshness and glottal attack. J. Speech Hear. Res., 1, 344-349 (1958).

22. Rees, M., Some variables affecting perceived harshness. J. Speech Hear. Res., 1, 155-168 (1958).

23. Rousey, C.L., Stuttering severity during prolonged spontaneous speech. J. Speech Hear. Res., 1, 40-47 (1958).

24. Shelton, Jr., R., Knox, A., Arndt, W., and Elbert, M., The relationship between nasality score values and oral and nasal sound pressure level. J. Speech Hear. Res., 10, 549-557 (1967).

25. Sherman, D. Correlation between defective articulation and nasality in cleft palate speech. Cleft Palate Journal, 7, 626-629 (1970).

26. Sherman, D., Reliability and utility of individual ratings of severity of audible characteristics of stuttering. J. Speech Hear. Dis., 20, 11-16 (1955).

27. Sherman, D., The merits of backward playing of connected speech in the scaling of voice quality disorders. J. Speech Hear. Dis., 19, 312-321 (1954).

28. Sherman, D., Usefulness of the mean in psychological scaling of cleft palate speech. Cleft Palate Journal, 7, 622-625 (1970).

29. Sherman,D., and Cullinan,W., Several procedures for scaling articulation. J. Speech Hear. Res., 3, 191-198 (1960).

30. Sherman, D., and McDermott, R., Individual ratings of severity of moments of stuttering. J. Speech Hear. Res., 1, 61-67 (1958).

31 Sherman, D., and Moodie, C., Four psychological scaling methods applied to articulation defectiveness. J. Speech Hear. Dis., 22, 698-706 (1957).

32. Sherman, D., and Morrison, S., Reliability of individual ratings of severity of defective articulation. J. Speech Hear. Dis., 20, 352-358 (1955).

33. Sherman, D., Shriner, T., and Silverman, F., Psychological scaling of language development of children. Iowa Academy of Science 72, 366-371 (1965).

34. Sherman, D., and Silverman, F., Three psychological scaling methods applied to language development. J. Speech Hear. Res., 11, 837-841 (1968).

35. Shriner, T., A comparison of selected measures with psychological scale values of language development. J. Speech Hear. Res., 10, 828-835 (1967).

36. Shriner, T. and Sherman, D., An equation for assessing language development. J. Speech Hear. Res., 10, 41-48 (1967).

37. Spriestersbach, D., Assessing nasal quality in cleft palate speech of children. J. Speech Hear. Dis., 20, 266-270 (1955).

38. Spriestersbach, D., and Powers, G., Nasality in isolated vowels and connected speech of cleft palate speakers. J. Speech Hear. Res., 2, 40-45 (1959).

39. Templin, M.D., and Darley, F.L., The Templin-Darley tests of articulation. Iowa City, Iowa: Bureau of Educational Research and Service, Extension Division, University of Iowa, 1960.

40. Torgerson, W., Theory and Methods of Scaling. New York: Wiley (1958).

41. Van Hattum, R., Articulation and nasality in cleft palate speakers. J. Speech Hear. Res., 1, 383-387 (1958).

42. Van Riper, C., and Irwin, J. Voice and Articulation. Englewood Cliffs, N.J.: Prentice-Hall,Inc. (1958).

43. Williams, D., Wark, M., and Minifie, F., Ratings of stuttering by audio, visual, and audiovisual cues. J. Speech Hear. Res., 6, 91-100 (1963).

44. Winitz, H., Articulatory Acquisition and Behavior. New York: Appleton-Century-Crofts (1969).

45. Wright, H., Reliability of evaluations during basic articulation and stimulation testing. J. Speech Hear. Dis., Monogr. 4, 19-27 (1954).

46. Young, M., Anchoring and sequence effects for the category scaling of stuttering severity. J. Speech Hear. Res., 13, 360-368 (1970).

47. Young, M., and Prather, E., Measuring severity of stuttering using short segments of speech. J. Speech Hear. Res., 5, 256-262 (1962).

APPENDIX A

POSSIBLE USES OF EQUAL-APPEARING INTERVAL SCALING IN PUBLIC SCHOOL SPEECH THERAPY PROGRAMS

Scaling rather than being limited to just the original screening program as described in the introduction, could also be utilized in the following areas: a. pupil progress; b. therapists effectiveness; c. training new therapists; and d. training classroom teachers.

a. Pupil Progress

The goal of speech therapy is to rehabilitate the individual with a speech defect so that his speech does not call attention to itself or interfere with communication. At present the most common ways for a therapist to evaluate the progress of a particular child is to (a) administer an articulation test and do a phonetic inventory or (b) listen to the child's conversational speech on a particular day and in some manner estimate how close this child's speech is to "normal". The weaknesses of evaluating progress by administering an Articulation test and doing a phonetic inventory are that, they do not reveal the effect of the defect on communication. In addition, a child who has been receiving speech therapy is well aware of the role of the speech therapist and the articulation test. Many children will correctly produce the sound during the so called "testing period" and then consistently misarticulate the sound in conversational speech. Being fully aware of this most therapists will attempt to evaluate progress through conversational speech. Using this as a tool for measuring progress, the therapist basically can only evaluate how close to "normal" is this child's speech. Since she only has the present speech sample, she has no real way of evaluating progress i.e., how far has he come from the first day he entered therapy. A far better way to evaluate progress would be through the use of scaling. The therapist records the original speech of the individual and assigns a number on an Equal Appearing Interval scale. When she is attempting to measure progress, she again records the conversational speech and assigns a number on an Equal Appearing Interval scale. Using this method the therapist now has two numbers which she can use to evaluate progress. The child whose original position was a 7 and now is assigned a 4 certainly has made more progress than another child who moves from a 5 to a 4.

b. Therapist Effectiveness

The same data can be used in measuring a therapist's effectiveness. A therapist who is working with many children who show little or no progress might reexamine her goals and methods. Thus, the therapist is not only evaluating pupil progress but also her effectiveness as a therapist.

c. Training New Therapists

The experience a beginning therapist has with respect to the severity of a speech problem is limited to the type and severity of the cases the training institution was able to provide, e.g., in many instances the "university sample" is quite different than the "public school sample". By using the Equal Appearing Interval method a training tape[1] could be constructed for a particular school system which would directly reflect the limits of severity within that system. This tape could then be used with all beginning therapists and thus help to insure that they will use the same yardstick as their more experienced co-professionals.

d. Training Classroom Teachers

In some public schools the speech therapist, rather than doing a speech survey, relies on the classroom teacher to refer children needing help. Diehl and Sinnett (7) studied the efficiency of teacher referrals in a school speech program. They report that the classroom teachers failed to identify 42.7% of the speech cases. The teachers missed almost 40% of the articulation cases. More recently Bucher and Canahl (2) investigated whether or not first grade teachers could reliably judge Articulation defectiveness in first grade students. The results of the study indicated that first grade teachers could reliably (.92) judge articulation defectiveness in first grade students. However, the results also indicated that there was less agreement between the first grade teachers than a group of speech therapists. In addition, in this particular study the first grade teachers consistently judged articulation defectiveness more severely than the therapists. Based upon these results one might expect that teachers making referrals for speech therapy, particularly in the primary grades, would over refer rather than under refer. This problem, whether it be over referral or under referral, could also be resolved through the use of speech tapes with classroom teachers.

In summary, psychological scaling seems to be a useful tool in a public school speech program in the areas of speech screening, measuring pupil progress, therapist evaluation, training new therapists and classroom teachers.

---

[1] Morrison (20) constructed a severity scale of articulation defectiveness to be used in training individuals to recognize the various levels of severity. To date the severity scale has only been used to train individuals who are going to participate in a research study which is utilizing a scaling method.